

Smelling out Code Clones: Clone Detection Tool Evaluation and Corresponding Challenges

Rachel Gauci
The University of Edinburgh
s1402642@sms.ed.ac.uk

March 3, 2015

Abstract

Software clones have been an active area of research for the past two decades. However, although numerous clone detection tools are now available, only a small fraction of the literature has focused on tool evaluation, and this is in fact still an open problem. This is mostly due to the fact that standard information retrieval metrics such as recall and precision require a priori knowledge of clones already in the system. Detection tools also typically have a large number of parameters which are difficult to fine-tune for optimal performance on a particular software system, and different outputs produced by different tools add to the complexity of comparing one tool to another. In this review, we further explore the reasons why tool evaluation is still an open challenge, and present the current tools and frameworks targeted at mitigating these problems, focusing on the current standard benchmarks used to evaluate modern clone detection tools, and also presenting a recent method aimed at finding optimal tool configurations.

The research work disclosed in this publication is funded by the MASTER it! Scholarship Scheme (Malta). The scholarship is part-financed by the European Union - European Social Fund (ESF) under Operational Programme II - Cohesion Policy 2007-2013, “Empowering People for More Jobs and a Better Quality of Life”.

1 Introduction and Motivation

Software clones are similar code fragments found in a single codebase, and are typically the result of developers’ “cut-copy-paste-adapt techniques” [15]. Clones have been identified as giving off a *bad smell* [5] in code, and are generally considered to be a bad programming practice. Code clones lead to bug propagation, since a bug in a single code fragment will be replicated in all clones of that fragment, causing an increase in both difficulties and expenses related to project maintenance. Code clones also unnecessarily bloat the size of a software system, causing a strain on resources, and it is therefore desirable to eliminate or at least limit the number of clones in a system.

Software clones emerged as a research area 20 years ago in 1994 [17] and, as can be seen from figure 1, the field has lately experienced a significant increase in interest, with more and more authors contributing to the literature in recent years. From figure 2, we see that research ranges across four sub-areas [17] with varying degrees of popularity: clone analysis, clone detection, clone management and tool evaluation. In particular we note that most of the studies have focused on clone detection and clone analysis, two sub-areas which are concerned with the development of tools and techniques for clone identification, and the analysis of clone traits and features, such as reasons for their existence and their effects on codebases and project maintenance. Research in clone detection has resulted in more than 70 [19] detection tools being currently in existence, however only 3% [17] of the literature has been dedicated to the evaluation of these tools.

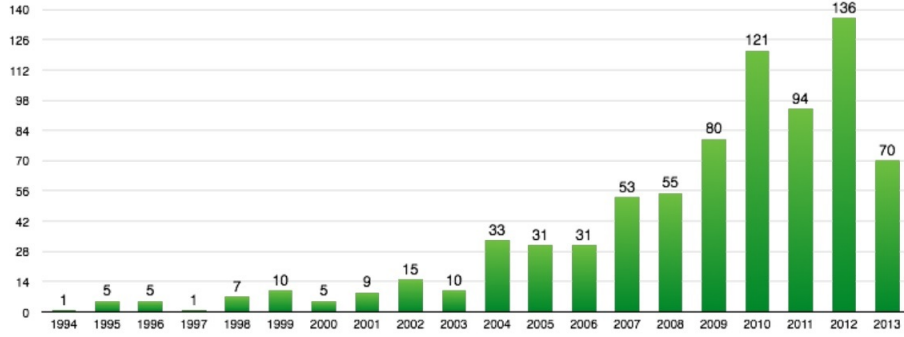


Figure 1: Yearly number of distinct authors contributing to clone research [17].

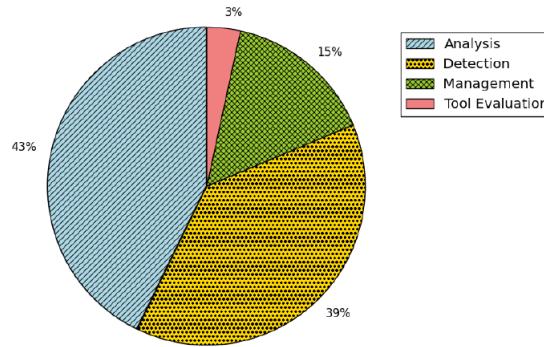


Figure 2: Proportion of publications in each category over the period 1994-2013 [17].

Tool evaluation is important for two main reasons. Firstly, when a new detection tool is made available, we would like to be able to evaluate its performance, both by itself and in relation to other detection tools; perhaps the new tool is better at identifying a particular type of clones, or has a lower false positive rate, or less demanding space and time requirements. Secondly, tool evaluation plays an important role in tool selection: given a particular software system, written in a particular programming language, having a particular size, and suspected to contain certain types of clones, we would like to know which detection tool would be the most suitable for the clone detection task, without having to try the tools out one by one and compare or verify their output.

Tool evaluation has however proved to be a very difficult task. To this day, evaluation still remains an open challenge [17], and in this review we explore the challenges in this sub-area of clone research. Currently, only two benchmarks exist for tool evaluation [17]: Bellon’s Framework [6] and Roy et al.’s Mutation & Injection Framework [16, 20], and this review focuses on their limitations while also presenting the recent EvaClone tool [21], which aims to mitigate the tool configuration problem threatening the validity of tool comparison experiments.

2 Clone Detection Tool Evaluation

2.1 Challenges in Tool Evaluation

Given a software system and a resultant set of clones detected by a particular tool, we would ideally be able to come up with quantitative measures for the tool’s coverage, whether it was able to find all the clones in the system, and the tool’s accuracy, whether any of the reported clones were actually

false positives. As mentioned in section §1, this kind of tool evaluation is not a trivial task, and this is due to three main reasons, which we outline below.

2.1.1 Lack of Reference Corpora

The standard metrics in the field of Information Retrieval, **recall** and **precision**, require a priori knowledge of the clones already in the system. Recall is concerned with a tool’s coverage. For a detection tool T and a software system S , recall can be calculated as in equation (1), where a recall value of 1 would indicate that the tool was able to detect all clones in the system.

$$\text{recall}_{T,S} = \frac{\text{number of detected clones}}{\text{total number of clones in system}} \quad (1)$$

Precision, on the other hand, is concerned with accuracy, and for a detection tool T and software system S , can be calculated as in equation (2), where a value of 1 would indicate that all reported clones were indeed true positives

$$\text{precision}_{T,S} = \frac{\text{number of actual clones from reported clones}}{\text{total number of reported clones}} \quad (2)$$

The two metrics work at odds with each other, since relaxing a tool’s similarity measure will increase its recall while sacrificing its precision, and vice-versa. It is easy, for example, to come up with a tool having 100% recall by simply reporting all possible combinations of code fragments as clones, but such a system would have a very low precision value and would, in effect, be useless. Therefore a good clone detection tool should have both high recall and high precision. The main problem with these metrics is, however, that they cannot be calculated without first identifying the clones in the system.

As pointed out by Bellon et al. [6], there are three obvious, naïve ways in which a reference corpus can be constructed from the candidate clones reported by a number of detection tools:

- (a) **union of candidates:** using this as a reference corpus, we would end up with a large number of false positives, as it is enough for one tool to classify two clone fragments as clones for the pair to be included in the reference corpus. Not only this, but this approach would result in a precision of 1 for all tools, which is not in any way representative of the tools’ true precisions, but merely a result of an inaccurate definition of what constitutes a reference clone.
- (b) **intersection of candidates found by all tools:** using the intersection instead of the union would have the opposite effect, leading to an inaccurate recall value of 1 for all tools, and a large number of true negatives, since failure of a single detection tool to identify an actual clone would lead to that clone being excluded from the reference corpus.
- (c) **intersection of candidates found by at least N Tools:** this approach is essentially a compromise between (a) and (b) above, but is still not a good solution, as no matter what value is chosen for N , N tools might still report a false positive, while $N - 1$ tools identify a true positive [6].

Bellon’s Framework [6] and the Mutation & Injection Framework [16,20] address the problems of a missing reference corpus by constructing a corpus of reference clones using two very different methodologies. In Bellon et al.’s approach [6], 2% of the 325,935 clones identified by 6 different clone detectors on 8 large C and Java software systems were manually “oracled” by Stefan Bellon, and the ones accepted by Bellon as actual clones were included as clones in the reference corpus. In Roy et al.’s approach [16,20], mutation operators are used to automatically generate synthetic clones from randomly-selected code fragments in the system, where the mutation operators aim to imitate the copy, paste and modify operations performed by developers which lead to real code clones in software systems. The artificial clones generated by the mutation operators are then injected into the software systems and used as the reference corpus for evaluation. We discuss the advantages and limitations of the approaches adapted by these two benchmarks in section 2.2.

2.1.2 The Tool Configuration Problem

Each detection tool has several different parameters and configuration options and, for optimal performance of the tool, these need to be fine-tuned according to the software system being analyzed. This problem is well recognized in clone literature, with as many as 113 papers commenting on the effects that tool configurations can have on empirical evaluations, 57 of which have a dedicated section discussing *threats to validity* [21]. This “confounding configuration choice problem”, as coined by Wang et al. [21], threatens the validity of evaluation results, since differences observed in tool performance might not be related to properties of the clones or the tools themselves, but may be the result of the configurations and parameters used in that particular experiment.

EvaClone¹ [21] aims to mitigate this confounding configuration problem. Given a set of detection tools and software systems, it makes use of a genetic algorithm to search the space of all possible tool configurations, in order to find the optimal configurations for each tool. The default fitness function aims to maximize tool agreement, i.e. maximize the size of the intersection of clone candidates, but the EvaClone framework allows for any choice of fitness function since the function itself is a parameter, and we further review this framework in section 2.3.

2.1.3 Incompatible Tool Output Formats

Different clone detectors output their results in different formats, such as HTML, XML and plain-text, and this complicates the process of head-to-head empirical evaluations. Not only this, but the clone results differ in terms of clone size and granularity, and the clones types identified by the tool and their degree of similarity. We do not focus on this problem in this review, since the evaluations carried out for Bellon’s Framework [6], the Mutation & Injection Framework [16, 20] and EvaClone [21] have managed to avert this problem by transforming and normalizing the outputs of the evaluated tools, but we acknowledge that it does indeed present a challenge to the evaluation of different tools.

Proposed formats aimed at standardizing tool output include Harder and Göde’s *Rich Clone Format* (RCF) [8] and Duala-Ekoko et al.’s Clone Region Descriptor (CRD) [7]. Kapsner et al. [11] also presented a draft proposal for the setup of a unified clone model, and set up an online wiki at <http://www.softwareclones.org/ucm> [12] for discussion and further proposals, although we noted that the discussion forums on the wiki are unfortunately inactive.

2.2 Limitations of Current Evaluation Benchmarks

Bellon’s Framework [6] consists of a set of reference clones constructed by manual verification of the clone candidates reported by 6 different tools which participated in an evaluation experiment in 2002. These tools represented the state-of-the-art at the time, including Dup [1, 2] and CloneDR [3, 4], and an early version of CCFinder [10], the current, most widely-cited detection tool in clone literature [15].

The outcome of the experiment presented several interesting results concerning the participating tools, including indications that token-based and text-based tools have higher recall, while tree-based tools have higher precision [6].² We note however that the experimental setup suffers from a number of limitations, and believe these should be taken into account when using the framework for evaluation. Our first concern is that the reference corpus was built by a single person, in this case Stefan Bellon, and this could have introduced unknown bias into the corpus. Different developers have different tolerance for clones, and as pointed out by Svajlenko et al. [20] and Wang et al. [21], the reliability of human judges is questionable, as even expert judges often disagree on whether two code fragments are cloned or not. Even in the case where human intervention is necessary, we suggest that bias can perhaps be eliminated or at least reduced by including more than one individual in the corpus-building process.

¹We use the term “EvaClone” to represent both the *EvaClone* and *CloudEvaClone* tools, where the latter is simply a cloud-based, parallelized version of the original *EvaClone* desktop application [21].

²A good explanation of these and other clone detection techniques can be found in Rattan et al.’s literature survey [15], together with examples of tools using each technique and an overview of empirical comparison studies.

Another limitation of Bellon’s framework is that only 2% of the clone candidates were “oracled” because of time limitations, as the 325,935 candidates took Bellon 77 hours to classify [6]. As a countermeasure, two evaluation experiments were conducted, one after each 1% of the candidates were classified, and the authors show that the distribution of the detected clones was similar after both experiments, where the number of candidate clones per tool approximately doubled after the second experiment.

Bellon et al.’s experiment [6] also involved the injection of 50 type 1, 2 and 3 clones into the software systems, but the detection tools were only able to locate 43 of them. This brings into question the reliability of using the tools’ results themselves as a reference corpus, since actual clones might go undetected and will therefore not be taken into account during the evaluation process. The corpus could also be biased towards the detection tools used to create it, putting other tools at a disadvantage, and the results of a recent study by Svajlenko et al. [19] do indeed seem to suggest that this is the case. This study [19] involved a comparison of the performance of two versions of CCFinder [9,10]: the one used in the original experiment in 2002 [10] and the more recent CCFinderX [9]. The results of the study [19] show that, according to Bellon’s Framework [6], the recall for the newer version is worse for almost all three types of clones on both Java and C systems. This brings to light what is possibly the current biggest limitation of Bellon’s Framework [6]: its corpus was built from a number of tools which were contemporary to 2002, and it might not be sufficient for evaluating the performance of modern tools [19].

Roy et al.’s Mutation & Injection Framework [16,20] avoids this problem by creating artificial clones from original subject systems through the use of mutation operators, and recent work [20] on the original framework [16] has generalized the framework for use with any clone detection tool. The automatic generation of clones also avoids the problem of human bias in the corpus generation process, and the framework was in fact the first attempt at a fully automated evaluation process [16]. However, as pointed out by Svajlenko et al. [19], this is also the framework’s major “threat to validity”, since the artificial clones generated through mutation analysis might not be representative of actual clones generated by developers.

For a more thorough comparison between these two frameworks, we refer the reader to Svajlenko et al.’s recent study [19] comparing the recall of 11 modern clone detection tools using both frameworks, and comparing them to expected recall values. The Mutation & Injection Framework [16] performs better overall [19], although we would have also liked to see precision measures included in the study since these metrics should be taken into account together, as discussed in section 2.1. We also note that the authors acknowledge the fact that their expectations, although well-researched, could contain inaccuracies, and we also add that they could contain an element of bias, since the authors of the study [19] are also the creators of the Mutation & Injection Framework [16,20].

2.3 Addressing The Tool Configuration Problem

As outlined in section 2.1, the tool configuration problem is well acknowledged but unfortunately not well addressed [21]. During Bellon et al.’s experiment [6], the evaluation was split into two parts: *mandatory*, using the tools’ default configurations, and *voluntary*, where the tools’ authors could optionally submit the tool with its configurations fine-tuned for the subject systems. However, only two of the authors submitted their tool for the voluntary run. In Svajlenko et al.’s recent study evaluating modern clone detection tools [19], the default configurations were used, and this is in fact included in the paper as another “threat to validity”.

Wang et al. [21] show that this is a recurring problem in clone literature, where a large proportion of the literature simply uses the default configuration settings for evaluation. Using the optimal tool configurations, as determined by EvaClone [21], the fitness values for different clone detectors on Java and C systems was improved by up to 21.9% and 10.6%, respectively, showing that tool configuration can have a significant effect on the tool’s performance, and is indeed something to be taken into consideration when carrying out comparative studies.

However, similar to what we have observed in other examples of clone literature, EvaClone also

comes with *threats to validity* [21]. Firstly, the subject systems used in the tool’s evaluation [21] were all open source, and the authors admit that the results might not be representative of other types of software systems. Secondly the tool’s fitness function presents another configuration problem in itself. The default fitness function aims to maximize the agreement between tools, and our main concern is that such a metric can inflate the recall value, as explained in section 2.1, since this is essentially another spin on using the intersection of clone candidates as a reference corpus [6]. In fact, when Wang et al. [21] used EvaClone’s optimal tool configurations to evaluate performance on the *psql* and *swing* systems using Bellon’s Framework [6], they found that their fitness function favours recall over precision, supporting our hypothesis that a different fitness function might be more appropriate. The work on EvaClone and corresponding results [21] however do confirm that empirical studies are potentially flawed in the way evaluation has been performed, and it would be very interesting to see how the results from Svajlenko et al.’s thorough study of modern tool evaluation [19] would be impacted through the use of optimal configurations.

3 Conclusion and Future Work

Through this review, we have highlighted the main challenges involved in the open problem of performing clone detection tool evaluation, and presented the current, state-of-the-art frameworks aimed at mitigating these challenges: Bellon’s Framework [6] and the Mutation & Injection Framework [16, 20], which have been used to generate reference corpora to enable the calculation of evaluation metrics, and the EvaClone framework, which can be used for automatic determination of optimal tool configurations, based on a pre-determined fitness function.

Despite these being the state-of-the-art, all frameworks have been shown to contain several limitations, but software clones are a very active research area [17], and we are therefore hopeful that further progress will be made in the sub-area of tool evaluation. Svajlenko et al. [19] have confirmed that their priorities for future work involves updating Bellon’s Framework [6] with clones detected by modern tools, while Roy et al. [17] suggest that significant progress could be made if research teams actually start exchanging data for reasons other than simply benchmarking their tools.

Clone detection is also being applied to areas other than software, such as clones in UML domain models [18] and MATLAB/Simulink models [13]. Other research, such as that by Rahman et al. [14], focuses on the advantages of having clones in a software system, and suggests that clones might not be a *bad smell* after all [14], although we note that this study also comes with its own *threats to validity* section, something which is unfortunately still very common in clone literature.

However, despite all these difficulties, clone management is gathering momentum in the industry. Several detection tools are now available as plug-ins for the Eclipse IDE, and clone management features have also been added to Microsoft Visual Studio [17], and so we are hopeful that, with increasing interest in both the research and industrial communities, new approaches to the problem of tool evaluation will be proposed, so that the current limitations can be addressed, enabling comparative, evaluation experiments to be carried out without any further threats to validity.

References

- [1] Brenda S Baker. On Finding Duplication and Near-Duplication in Large Software Systems. In *Reverse Engineering, 1995., Proceedings of 2nd Working Conference on*, pages 86–95. IEEE, 1995.
- [2] Brenda S Baker. Parameterized Duplication in Strings: Algorithms and an Application to Software Maintenance. *SIAM Journal on Computing*, 26(5):1343–1362, 1997.
- [3] Ira D Baxter, Christopher Pidgeon, and Michael Mehlich. DMS®: Program Transformations for Practical Scalable Software Evolution. In *Proceedings of the 26th International Conference on Software Engineering*, pages 625–634. IEEE Computer Society, 2004.

- [4] Ira D Baxter, Andrew Yahin, Leonardo Moura, Marcelo Sant’Anna, and Lorraine Bier. Clone Detection Using Abstract Syntax Trees. In *Software Maintenance, 1998. Proceedings., International Conference on*, pages 368–377. IEEE, 1998.
- [5] Kent Beck, Martin Fowler, and Grandma Beck. Bad smells in code. *Refactoring: Improving the design of existing code*, pages 75–88, 1999.
- [6] Stefan Bellon, Rainer Koschke, Giuliano Antoniol, Jens Krinke, and Ettore Merlo. Comparison and Evaluation of Clone Detection Tools. *Software Engineering, IEEE Transactions on*, 33(9):577–591, 2007.
- [7] Ekwa Duala-Ekoko and Martin P Robillard. Clone Region Descriptors: Representing and Tracking Duplication in Source Code. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, 20(1):3, 2010.
- [8] Jan Harder and Nils Göde. Efficiently Handling Clone Data: RCF and cyclone. In *Proceedings of the 5th International Workshop on Software Clones*, pages 81–82. ACM, 2011.
- [9] Toshihiro Kamiya. The Official CCFinderX website. <http://www.ccfinder.net/ccfinderx.html>, 2008. [Online; accessed 27-Nov-2014].
- [10] Toshihiro Kamiya, Shinji Kusumoto, and Katsuro Inoue. CCFinder: A Multilinguistic Token-Based Code Clone Detection System for Large Scale Source Code. *Software Engineering, IEEE Transactions on*, 28(7):654–670, 2002.
- [11] CJ Kapser, Jan Harder, and Ira Baxter. A Common Conceptual Model for Clone Detection Results. In *Software Clones (IWSC), 2012 6th International Workshop on*, pages 72–73. IEEE, 2012.
- [12] CJ Kapser, Jan Harder, and Ira Baxter. Unified Clone Model Wiki. <http://www.ccfinder.net/ccfinderx.html>, 2012. [Online; accessed 28-Nov-2014].
- [13] Nam H Pham, Hoan Anh Nguyen, Tung Thanh Nguyen, Jafar M Al-Kofahi, and Tien N Nguyen. Complete and Accurate Clone Detection in Graph-Based Models. In *Proceedings of the 31st International Conference on Software Engineering*, pages 276–286. IEEE Computer Society, 2009.
- [14] Foyzur Rahman, Christian Bird, and Premkumar Devanbu. Clones: What is that smell? *Empirical Software Engineering*, 17(4-5):503–530, 2012.
- [15] Dhavleesh Rattan, Rajesh Bhatia, and Maninder Singh. Software clone detection: A systematic review. *Information and Software Technology*, 55(7):1165–1199, 2013.
- [16] Chanchal K Roy and James R Cordy. A Mutation / Injection-based Automatic Framework for Evaluating Code Clone Detection Tools. In *Software Testing, Verification and Validation Workshops, 2009. ICSTW’09. International Conference on*, pages 157–166. IEEE, 2009.
- [17] Chanchal K Roy, Minhaz F Zibran, and Rainer Koschke. The Vision of Software Clone Management: Past, Present, and Future (Keynote Paper). In *Software Maintenance, Reengineering and Reverse Engineering (CSMR-WCRE), 2014 Software Evolution Week-IEEE Conference on*, pages 18–33. IEEE, 2014.
- [18] Harald Störkle. Towards Clone Detection in UML Domain Models. In *Proceedings of the Fourth European Conference on Software Architecture: Companion Volume*, pages 285–293. ACM, 2010.
- [19] Jeffrey Svajlenko and Chanchal K Roy. Evaluating Modern Clone Detection Tools. *Proc. ICSME*, 2014.

- [20] Jeffrey Svajlenko, Chanchal K Roy, and James R Cordy. A Mutation Analysis Based Benchmarking Framework for Clone Detectors. In *Software Clones (IWSC), 2013 7th International Workshop on*, pages 8–9. IEEE, 2013.
- [21] Tiantian Wang, Mark Harman, Yue Jia, and Jens Krinke. Searching for Better Configurations: A Rigorous Approach to Clone Evaluation. In *Proceedings of the 2013 9th Joint Meeting on Foundations of Software Engineering*, pages 455–465. ACM, 2013.